Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/20BE09Krickovic

SISTEM ZA PREPORUKU VESTI UPOTREBOM PERSONALIZOVANIH KRATKOTRAJNIH REPREZENTACIJA

NEWS RECOMMENDATION SYSTEM USING PERSONALIZED SHORT TERM REPRESENTATIONS

Luka Kričković, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *Pojavom internet portala koji ne prikazuju samo svoje autorske članke, nego agregiraju članke sa više izvora, sistemi za preporuku vesti su postali neophodni, kako bi korisnici došli do sadržaja koji ih zanima. Takođe, ovakvi sistemi se mogu primeniti i na drugim platformama poput Twitter-a, gde preporuka tekstualnog sadržaja direktno dovodi do kvalitetnijeg korišćenja platforme. U ovom radu će biti opisan sistem za preporuku tekstualnog sadržaja, a prvenstveno vesti, koji ume da razlikuje korisnikove kratkotrajne i dugotrajne sklonosti, pa na osnovu istorijata pregleda članaka generiše preporuke. Rezultati ovog modela nad MIND-demo skupom podataka su 64.2 AUC, 29,43 mean MRR, 38.78 NDCG@5 i 32.37 NDCG@10.*

Ključne reči: *preporuka vesti, mašinsko učenje, duboko učenje*

Abstract – *With the emergence of internet news portals, especially ones that publish their articles and aggregate articles from other sources, news recommendation systems have become a growing part of the industry. On the other hand, social media websites like Twitter can also benefit from sound textual recommendation systems, as the quality of users' interaction directly depends on the content they serve. This paper aims to introduce an implementation of a news recommendation model that differentiates between users' short- and long-term interests and, with them in mind, generates recommendations for the users. The model we will describe has achieved 64.2 AUC, 29,43 mean MRR, 38.78 NDCG@5, and 32.37 NDCG@10.*

Keywords: *news recommendation, machine learning, deep learning*

1. UVOD

Popularizacijom mobilnih telefona i drugih prenosivih uređaja, štampane vesti su sve manje relevantne čitaocima, a samim time i novinskim kućama. Povećanjem popularnosti internet portala, nastao je novi tip novinskih sajtova, koji ne objavljuju samo svoje autorske članke, već sakupljaju i članke sa drugih izvora i serviraju korisnicima personalizovan sadržaj.

Kako bi servirali korisnicima relevantan sadržaj, ove platforme moraju imati kvalitetne sisteme za preporuku sadržaja. Primer ovakvih platformi je *Microsoft news* [1]. Sa druge strane, ovakvi sistemi se mogu primeniti in a blog sajtove poput *Medium*-a [2], koji takođe koristi sličan sistem kako bi korisnicima servirao sadržaj, tako da se kvalitet ovakvog sistema direktno odražava na kvalitet interakcije korisnika platforme.

Prvi sistemi za generisanje preporuka su koristili *content filtering* tehnike, kako bi razumele sadržaj. Drugim rečima, ovi sistemi su analizirali sadržaj, ali ne i čoveka kome preporučuju sadržaj. Kreirani su profili za interesantne sadržaje i na osnovu toga su se klasifikovali binarno: preporučiti ili ne. Potom su usledile *collaborative filtering* tehnike, koje su značajne jer su počele da uzimaju u obzir i korisnike kojima preporučuju sadržaj. Novi problem koji nastaje upotrebom ovih tehnika je upravljanje velikom količinom podataka [3]. Kako je sve više sadržaja dostupno na internetu, a i dinamika konzumiranja sadržaja postaje sve kompleksnija, nova rešenja su postala potrebna kako bi korisnici zaista dobijali dobre preporuke.

Moderna rešenja za preporuku tekstualnog sadržaja, a pogotovo vesti, su najbolje opisana u radu [4]. Sva moderna rešenja, na kojima se bazira model opisan u ovom radu, se temelje na dubokom učenju i na arhitekturama sličnim transformer mrežama. Prvi rad koji predlaže moderniju arhitekturu za sisteme za preporuku vesti je DKN (eng. *Deep Knowledge-aware Network*) [5], koji predlaže da se sistem podeli u tri komponente, enkoder članaka, enkoder korisnika i podsistem za generisanje preporuka.

Rešenje opisano u ovom radu se sastoji iz sva tri podsistema koje predlaže rad [5], ali su sve pojedinačne komponente izmenjene, kako bi drugačije analizirali članke i korisnike i time generisali kvalitetnije i više personalizovane preporuke.

Sistem predstavljen u radu ume da razlikuje korisnikove kratkotrajne i dugotrajne interese, tako da određene članke preporučuje samo u odgovarajućem vremenskom periodu, a putem enkodovanja naslova članaka ume da prepozna koje će reči privući pažnju određenom korisniku.

2. PRETHODNA REŠENJA

Stariji sistemi za preporuku su koristili *content* i *collaborative filtering*, ali u ovom poglavlju će veći fokus

---

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.

biti na modele koji koriste enkoderske arhitekture za generisanje preporuka. Najpopularnija rešenja ovog tipa su prikazana u radu [4].

Prvi popularniji sistem koji se oslanja na enkodersku arhitekturu je bio DKN [5], koji je predlagao podelu sistema na enkoder članaka, enkoder korisnika i podsistem za generisanje preporuka. Enkoder članaka prvo uzima naslov članka i vektorizuje ga upotrebom *knowledge graph* tehnika, kako bi vektori sadržali kontekst iza reči u naslova [5]. Nakon inicijalne vektorizacije, članci prolaze kroz konvolutivnu neuronsku mrežu, koja priprema vektore za *attention* mehanizam koji sledi. Ovaj mehanizam služi da izdvoji ključne, odnosno najbitnije reči iz naslova članaka. Enkoder korisnika podrazumeva iteriranje kroz istorijat članaka i vektorizaciju svakog naslova teksta. Kada se svi tekstovi vektorizuju, novonastali vektori se sumiraju i šalju u klasifikator, zajedno sa vektorskom reprezentacijom kandidatskog članka. Klasifikator je u ovom slučaju sekvencijalna mreža, koja određuje da li treba korisniku preporučiti članak ili ne. Takođe, ovaj rad uvodi i AUC (eng. *Area Under Curve*) metriku, kojom se evaluiraju svi moderni radovi na ovu temu. Ovaj model postiže 65.9% AUC nad skupom podataka kojeg su autori pripremili preuzimanjem podataka sa različitih portala. Za poređenje, najbolji prethodni model, LibFM, je postizao 59.7% AUC.

Jedan od zanimljivijih radova na ovu temu bio je [6], jer uvodi personalizovani enkoder članaka, koji vektorizuje članke na način da naglasi reči iz naslova koje pojedinačnom korisniku mogu privući pažnju, za razliku od ostalih primera koji su se oslanjali na upotrebu generičkih *attention* mehanizama. Rezultati koje ovaj model (u daljem tekstu NPA model) postiže su 62.45% AUC metrike, što je više od DKN, koji je postizao 58.63% nad istim skupom podataka.

Prvi rad koji se više fokusira na enkoder korisnika nego na enkoder članka bio je [7], koji predstavlja LSTUR model. Ovaj enkoder je specifičan jer ne koristi samo jednu vektorsku reprezentaciju korisnika, nego kreira dve: kratkotrajnu i dugotrajnu reprezentaciju. Dugotrajni interesi svakog korisnika se čuvaju u obliku *lookup* tabele, a kratkotrajni interesi se modeluju pomoću enkodera članaka, čiji se izlazi šalju u rekurentnu neuronsku mrežu. Rekurentne mreže su spojene u lanac, tako da izlaz iz rekurentne jedinice inicijalizuje sledeću jedinicu. Ukoliko je poslednja rekurentna jedinica, odnosno poslednji članak, izlaz se šalje direktno u klasifikator, koji je ponovo sekvencijalna mreža. Postoje dve verzije ove arhitekture, inicijalizaciona i konkatenaciona, gde inicijalizaciona koristi dugotrajne reprezentacije korisnika za inicijalizaciju prve rekurentne jedinice, dok konkatenaciona konkatenira vektorsku reprezentaciju korisnika na izlaz iz poslednje rekurentne jedinice.

3. SKUPOVI PODATAKA

Kako je inspiracija za implementaciju ovog sistema proistekla iz MIND takmičenja, jedini podesan skup za ovaj model bio je upravo MIND (eng. *Microsoft News Dataset*). Ovaj skup podataka je specifičan, jer je prvi javno dostupan skup podataka koji, osim podataka o člancima, sadrži i podatke o utiscima korisnika, kao i o sesijama korisnika koji interaguju sa *Microsoft News* platformom [1].

Mind skup podataka se sastoji iz 160.000 članaka na engleskom jeziku, kao i 15 miliona utisaka od milion korisnika *Microsoft News* platforme. Sastoji se iz četiri datoteke, „behaviours.tsv“, „news.tsv“, „entityembedding.vec“ i „relationembedding.vec“. *Behaviours* datoteka sadrži podatke o sesijama korisnika, modelovane kroz istorijat članaka koje je korisnik pregledao, uz dodatni podatak o vremenu početka sesije i anonimizovanom identifikatoru korisnika. Sesije ne sadrže samo samo članke koje je korisnik kliknuo, nego sadrže i određen broj članaka koje korisnik nije kliknuo, ali su mu servirani. *News* dokument sadrži tekstove i naslove vesti, ali sadrži i meta podatke o člancima, poput kategorije, pod kategorije, *url*-a, apstrakta, kao i embedinga, koji se koristi za inicijalizaciju embeding slojeva.

Behaviours datoteka sadrži utiske korisnika, modelovane kroz identifikatore utiska i korisnika čiji je utisak, vremena kreiranja utiska, istorijata članaka koje je korisnik pregledao, skupa naslova koji su predstavljeni korisniku anotirani sa 1 ili 0 u zavisnosti od toga da li je korisnik kliknuo na naslov ili ne. *News* dokument sadrži podatke o člancima poput identifikatora članka, kategorije, pod kategorije, naslova, apstrakta, linka ka članku, kao i podatke poput embedinga, naslova i apstrakta. *Entity* i *relation embedding* datoteke sadrže embeding vektore entiteta i relacija između entiteta. Za potrebe ovog projekta je korišćena demo verzija ovog skupa, kako bi se istrenirao model u prihvatljivom vremenskom periodu. Ovaj skup je prerađen tako što je vektorizovan upotrebom *word2vec* algoritma. Skup podataka je podeljen na trening, test i validacioni podskup.

Zbog postojanja *Microsoft*-ovog *recommenders* API-a, nisu ručno preuzimani skupovi i deljeni u podskupove, nego su pozivom metode „download_deeprec_resources“ preuzeti „MINDdemo_train“ i „MINDdemo_dev“, od kojih prvi služi za trening i test, a drugi za validaciju.

4. METODOLOGIJA

U ovom poglavlju će biti opisana arhitektura modela za preporuku vesti kroz tri poglavlja: (1) enkoder članka, (2) enkoder korisnika, (3) podsistem za generisanje preporuke. Takođe, u odvojenom poglavlju će biti opisan proces obučavanja modela.

4.1. Enkoder članaka

Ulaz u enkoder članaka je tekst naslova članka, koji ne prolazi kroz pretprocesiranje, jer se većina logike procesiranja odvija u okviru modela. Prvi sloj enkodera je vektorizator, koji od svake reči kreira embeding vektor sa *n* elemenata, gde je *n* u ovom slučaju 200 (hiper-parametar kojeg propisuje MIND takmičenje). Sledeći korak je konvolutivna neuronska mreža, koja priprema vektore za *attention* mehanizam. Svrha upotrebe konvolutivnih mreža u obradi tekstova je izvlačenje naglašenih reči iz rečenica, tako da dalji mehanizam

preporuke ne tretira sve reči ravnopravno. Na primer, model opisan u ovom radu, kao i model iz rada [6], naglašava reč "NBA" iz rečenice „NBA season is about to begin", ali je zanimljivo što naglašava i reči poput „Shock", „Disbelief" i slično, koje su karakteristika takozvanih *clickbait* naslova, koje su svakako i upotrebljene kako bi privukle pažnju čitaocima. Nakon pripreme vektora kroz konvolutivnu mrežu, sledeći korak u obradi je prolaz kroz *attentive pooling* mehanizam, koji naglašava samo reči koje su u fokusu rečenice. Ovaj mehanizam je prilično sličan konvolutivnoj mreži po svojoj svrsi, ali za razliku od konvolutivne mreže, ne naglašavaju se reči poput „Shock" i „Disbelief", nego se naglašavaju reči koje su semantički značajnije za rečenicu. Na prethodnom primeru, iz „NBA season is about to begin" bi *attentive pooling* mehanizam naglasio „NBA" i „begin", jer nose najveće semantičko značenje u rečenici.

4.2. Enkoder korisnika

Funkcija enkodera korisnika je da modeluje osobu koja interaguje sa sistemom kroz njihov istorijat pretrage. Ulaz u ovaj sistem je korisnikov identifikator, kao i njegov istorijat pročitanih članaka. Specifično za ovaj enkoder korisnika jeste da modeluje korisnikove kratkotrajne i dugotrajne interese. Na primer, ukoliko korisnik pogleda članak o formuli 1, verovatno će ga konstantno interesovati članci o vozačima i povezanim sportskim događajima. Nasuprot ovome, ukoliko korisnik pročita vest o filmu *Top Gun,* verovatno će ga u kratkom vremenskom periodu interesovati članci o *Tom Cruise*-u i drugim glumcima iz filma, ali ga to neće zanimati konstantno, nego do određenog vremenskog praga. Ovo se postiže izdvajanjem dugotrajnih interesa korisnika u *lookup* tabelu, u ključ-vrednost formatu, gde je ključ embedovan identifikator korisnika, a vrednost niz embedovanih naslova članaka. Ova tabela se kreira prilikom inicijalizacije modela, za korisnike koji su registrovani u sistemu, analizom istorijata preporuke i izdvajanjem grupa članaka koje korisnik konstantno otvara, dok se kratkotrajni interesi modeluju za vreme generisanja preporuke.

Svi članci, koji se smatraju kratkotrajnim interesima korisnika, se provlače kroz enkoder članaka opisan u prethodnom poglavlju. Vektorske reprezentacije članaka se prosleđuju svaka u svoju rekurentnu neuronsku mrežu (GRU, eng. *Gated Recurrent Unit*). Ove rekurentne mreže su povezane u lanac, tako da izlaz iz jedne mreže predstavlja težine inicijalizacije sledeće mreže. Dolazi se do pitanja: (1) kako inicijalizovati prvu rekurentnu mrežu i (2) kako upotrebiti podatke o dugotrajnim interesima korisnika?

Odgovor na ova pitanja pruža rad [7] i se može dati kroz dva pristupa: (1) inicijalizacioni i (2) konkatenacioni. Po prvom pristupu, vrednost *i* iz *lookup* tabele, zajedno sa embedovanim identifikatorom korisnika se koriste za inicijalizaciju težina prve rekurentne mreže, dok se po drugom pristupu izlaz iz poslednje rekurentne mreže konkatenira na vrednosti iz *lookup* tabele i tako šalje na podsistem za generisanje preporuke. Ne postoje jasne prednosti i jednog od ova pristupa, ali u okviru eksperimenata sprovedenih u ovom radu, neznatno bolje rezultate je konzistentno postizala arhitektura sa inicijalizacionim pristupom, pod pretpostavkom da su težine poslate na ulaz prve rekurentne jedinice ipak optimizovane procesom obučavanja modela. Iako je bolje rezultate postizao inicijalizacioni pristup, ovi rezultati nisu toliko drastično bolji da bi se moglo garantovati da je bolji od konkatenacionog.

4.3. Podsistem za generisanje preporuke

Ovaj podsistem je najjednostavniji od navedena tri i služi za obradu rezultata iz enkodera vesti i korisnika. Ulaz u ovaj sistem su generisane vektorske reprezentacije kandidatskog članka i korisnika, gde ovi vektori prolaze kroz dva odvojena „pod modela", čiji se rezultati agregiraju. Prvi pod model koristi običan skalarni proizvod, čime množi dva embedinga i rezultat provlači kroz *softmax*. Drugi pod model koristi *Time Distributed* sloj za modelovanje kandidatskog članka, koji se šalje na skalarni proizvod sa vektorskom reprezentacijom korisnika i koristi se sigmoidna aktivaciona funkcija. Na osnovu ova dva vektora se generiše preporuka upotrebom *recommenders* API-a [8].

4.4. Obučavanje modela

Kako se MIND skup podataka preuzima upotrebom *recommenders* API-a [8], dolazi unapred podeljen na tri podskupa, trening test i validacioni. U ovom radu, korišćena je demo verzija MIND skupa, jer su veće verzije previše računarski skupe za potrebe ovih eksperimenata. Podela na podskupove nije određena po procentima kao što se često radi (npr. 80%/10%/10%), nego je određena tako što su se podaci prikupljali kroz vremenski period od par nedelja, gde je šest dana poslednje nedelje upotrebljeno za test skup, a jedan dan poslednje nedelje upotrebljen za validacioni skup. Za demo verziju ovog skupa je samo skaliran broj članaka, tako da sadrži minimalan podskup originalnog skupa potreban za postizanje dobrih rezultata.

5. EVALUACIJA REŠENJA I REZULTATI

Za evaluaciju modela upotrebljene su sledeće metrike: (1) AUC (eng. *Area Under roc Curve*), *Mean* MRR, NDCG@5 i NDCG@10. Glavna od navedenih metrika je AUC i ona predstavlja površinu pokrivenu ROC (eng. *Receiver Operating Characteristic*) krivom. Ova kriva predstavlja grafički prikaz odnosa TPR (eng. *True Positive Rate*) i FPR (eng. *False Positive Rate*), koji se računaju na osnovu sledećih formula:

$$TPR = \frac{TP}{TP + FN},$$

$$FPR = \frac{FP}{FP + TN},$$

Gde TP predstavlja broj tačnih pozitivnih predikcija, FN broj lažnih negativnih predikcija, FP broj lažno pozitivnih i TN broj tačno negativnih predikcija. Sam AUC se dobija računanjem vrednosti integrala ove krive. Optimizacijom ove metrike se model tera da generiše tačne pozitivne preporuke, a uči se da zanemaruje nasumično tačne i tačno negativne preporuke [9].

Poređenje rezultata modela opisanog u ovom radu i modela na osnovu kojih je ovaj model baziran je dat na tabeli 1.

Tabela 1. *AUC, Mean MRR i NDCG rezultati modela*

Naziv modela	AUC	Mean MRR	NDCG@5	NDCG@10
Model opisan u ovom radu – demo skup	**64.2**	**29.43**	**38.78**	**32.37**
LSTUR – demo skup	52.01	22.14	22.92	29.12
LSTUR	**67.73**	**32.77**	**35.59**	**41.34**
NPA	66.69	32.24	34.94	40.68

Naglašena su dva reda, jer predstavljeni modeli nisu obučavani nad istim skupovima podataka. Kao što se može videti, model opisan u ovom radu postiže značajno bolje rezultate u odnosu na LSTUR [7], jer koristi personalizovani *attentive pooling*, za razliku od LSTUR modela koji koristi "generički" *attention* mehanizam. Takođe, kada se uporedi sa LSTUR i NPA modelima obučenim nad celim MIND skupom podataka, na osnovu rezultata iz rada [3], može se zaključiti da model opisan u ovom radu postiže gotovo iste rezultate kao modeli na kojima je baziran, iako je obučen nad drastično manjim skupom podataka.

6. ZAKLJUČAK

U ovom radu je predstavljeno rešenje za problem preporuke vesti, koje uzima u obzir korisnikove podatke prilikom generisanja vektorske reprezentacije naslova članaka i uspešno modeluje razliku između korisnikovih dugotrajnih i kratkotrajnih interesa.

Model se sastoji iz tri pod elementa, enkoder članaka, enkoder korisnika i podsistem za preporuku vesti. Enkoder članka koristi embeding sloj za generisanje embedinga svake pojedinačne reči iz naslova članka, nakon čega rezultujući vektori prolaze kroz konvolutivnu neuronsku mrežu i kroz personalizovani *attentive pooling* mehanizam.

Enkoder korisnika provlači sve članke iz istorije pretrage korisnika kroz enkoder članaka, nakon čega ih šalje u lanac rekurentnih mreža koje su inicijalizovane embedingom korisnikovih dugotrajnih interesa. Izlaz iz poslednje rekurentne jedinice u lancu se dalje šalje podsistemu za generisanje preporuke. Podsistem za generisanje preporuke kao ulaz, osim izlaza poslednje rekurentne jedinice iz enkodera korisnika, prima i vektorsku reprezentaciju članka. Za ova dva vektora se računa skalarni proizvod i na osnovu njega se generiše preporuka.

Moguća unapređenja za ovaj model su upotreba *multi head attention* mehanizma, poput onoga koji se pojavljuje u BERT modelu. Takođe, moguće je upotrebljavati BERT-ov tokenizator ili GPT tokenizator, koji kreiraju preciznije embedinge pojedinačnih reči. Poboljšanje bi se postiglo i korišćenjem ansambl modela, podelom skupa na *n* podskupova i nad svakim od podskupova obučiti po jednu instancu identičnog modela. Prilikom izvršavanja preporuke bi se računala prosečna vrednost preporuka svih *n* modela i time bi se postizale stabilnije performanse.

7. LITERATURA

[1] Microsoft news https://news.microsoft.com/ [datum pristupa 26.06.2022.]

[2] Medium https://www.medium.com [datum pristupa 26.06.2022.]

[3] Važnost sistema za preporuku https://medium.com/@Commons/the-importance-of-recommender-systems-36f86f92181 [datum pristupa 26.06.2022.]

[4] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation . In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* , pages 3597–3606, Online. Association for Computational Linguistics.

[5] Wang, Hongwei & Zhang, Fuzheng & Xie, Xing & Guo, Minyi. (2018). DKN: Deep Knowledge-Aware Network for News Recommendation. WWW '18: Proceedings of the 2018 World Wide Web Conference. 1835-1844. 10.1145/3178876.3186175.

[6] Wu, Chuhan & Wu, Fangzhao & An, Mingxiao & Huang, Jianqiang & Huang, Yongfeng & Xie, Xing. (2019). NPA: Neural News Recommendation with Personalized Attention.

[7] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In ACL, pages 336–345.

[8] Recommenders zvanična dokumentacija https://readthedocs.org/projects/microsoft-recommenders/ [datum pristupa: 16.7.2022.]

[9] Objašnjenje AUC metrike https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc [datum pristupa 16.7.2022/]

Kratka biografija:

Luka Kričković rođen je 10.2.1999. godine u Novom Sadu. Master rad na Fakultetu Tehničkih Nauka u Novom Sadu je odbranio 2022. godine.

kontakt: krickovicluka@gmail.com